Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: 2-meter Temperatures

Renate Hagedorn[2], Thomas M. Hamill[1] and Jeffrey S. Whitaker[1]

[1]*NOAA Earth System Research Laboratory, Boulder, Colorado*

[2]*European Centre for Medium-Range Weather Forecasts, Reading, England*

Submitted to *Monthly Weather Review*

9 October 2007

Corresponding Author Address:

Dr. Thomas M. Hamill
NOAA Earth System Research Lab, Physical Sciences Division
R/PSD1
325 Broadway
Boulder, CO 80303
Tom.Hamill@noaa.gov
Phone: (303) 497-3060 Fax (303) 497-6449

ABSTRACT

Recently, the European Centre for Medium-Range Weather Forecasts (ECMWF) produced a reforecast data set for a 2005 version of their ensemble forecast system. The data set consisted of 15-member reforecasts conducted for the 20-year period 1982-2001, with reforecasts computed once weekly from 1 September – 1 December. This data set was less robust than the daily reforecast data set produced for the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS), but it utilized a much higher-resolution, more recent model. This manuscript considers the calibration of 2-meter temperature forecasts using these reforecast data sets as well as samples of the last 30 days of training data. Nonhomogeneous Gaussian regression was used to calibrate forecasts at stations distributed across much of North America. It was observed that: (1) though the "raw" GFS forecasts (probabilities estimated from ensemble relative frequency) were commonly unskillful measured in continuous ranked probability skill score (*CRPSS*), after calibration with a 20-year set of weekly reforecasts, their skill exceeded those of the raw ECMWF forecasts. (2) Statistical calibration using the 20-year weekly ECMWF reforecast data set produced a large improvement relative to the raw ECMWF forecasts; the ~ 4-5 day calibrated reforecast-based product had a *CRPSS* as large as a 1-day raw forecast. (3) A calibrated multi-model GFS / ECMWF forecast trained on 20-year weekly reforecasts was slightly more skillful than either the individual calibrated GFS or ECMWF reforecast products. (4) Approximately 60-80 percent of the improvement from calibration resulted from the simple correction of time-averaged bias. (5) Improvements were generally larger at locations where the forecast skill was

originally lower, and these locations were commonly found in regions of complex terrain. (6) The past 30 days of forecasts were adequate as a training data set for short-lead forecasts, but longer-lead forecasts benefited from more training data. (7) A small but consistent improvement was produced by calibrating GFS forecasts using the full 25-year, daily reforecast training data set versus the subsampled, 20-year weekly training data set.

1. **Introduction**

A series of recent articles have introduced the use of reforecasts for the calibration of a variety of probabilistic weather-climate forecast problems, from week-2 forecasts (Hamill et al. 2004; Whitaker et al. 2006) to short-range precipitation forecast calibration (Hamill et al. 2006, Hamill and Whitaker 2006) to forecasts of approximately normally distributed fields such as geopotential and temperature (Wilks and Hamill 2007, Hamill and Whitaker 2007) to streamflow predictions (Clark and Hay 2004). The reforecast data set used was a reduced resolution, T62, 28-level, circa-1998 version of the Global Forecast System (GFS) from the National Centers for Environmental Prediction. Fifteen-member forecasts were available to 15 days lead for every day from 1979 to current. With a stable data assimilation and forecast system, the systematic errors of the forecast could be readily diagnosed and corrected. Calibration using reforecasts were able to adjust the forecasts to achieve substantial improvements in the skill and reliability of the forecasts, commonly to levels competitive with or exceeding those achieved by current-generation ensemble forecast systems without calibration.

The GFS model version used in these reforecast studies is now ~10 years out of date, and the reforecasts and real-time forecasts from it are run at a resolution far less than that used currently at operational weather prediction centers. Arguably, the dramatic improvement from the use of reforecasts may be due in large part to the substantial deficiencies of this forecast modeling system. Would the calibration of a modern-generation ensemble forecast system similarly benefit from the use of reforecasts?

Recently, the European Centre for Medium-Range Weather Forecasts (ECMWF) produced a more limited reforecast data set with a model version that was operational in the last half of 2005. They produced a 15-member reforecast once weekly from 1 September to 1 December, over a 20-year period from 1982 to 2001. Each forecast was run to 10 days lead using a T255, 40-level version of the ECMWF global forecast model. During the past decade, ECMWF global ensemble forecasts have consistently been the most skillful of those produced at any national center (e.g., Buizza et al. 2005), so calibration experiments with this model may be representative of the results that other centers may obtain with reforecasts over the next 5 years or so.

This data set allows us to ask and answer questions about reforecasts that were not possible with only the GFS data set. Some relevant questions include: (1) how does an old GFS model forecast that has been statistically adjusted with reforecasts compare with a probabilistic forecast estimated directly from the state-of-the-art ECMWF ensemble forecast system? (2) If this state-of-the-art system could also be calibrated using its own reforecast, would there still be substantial benefits from the calibration, or would they be much diminished relative to the improvement obtained with the older GFS forecast model? (3) Is a calibrated, multi-model combination more skillful than that provided solely by the ECMWF system? (4) How much of the benefits of calibration in a state-of-the-art model can be obtained using only a short time series of past forecasts and observations?

This article will consider the problem of the calibration of probabilistic calibration of 2-meter temperature forecasts. A companion article (Hamill et al. 2007) will discuss the calibration of 12-hourly accumulated precipitation forecasts. The calibration problems for each are unique; as will be shown, temperature forecasts tend to have more Gaussian errors and substantial improvements can be obtained with relatively modest training data sets. Calibration of non-normally distributed precipitation is more difficult, and larger samples tend to be needed to calibrate the more rare events.

Below, section 2 will review the data sets used in this experiment. Section 3 describes the calibration methodology and the methods for evaluating forecast skill. Section 4 provides results, and section 5 provides conclusions.

2. **Forecast and observational data sets used**.

a. ECMWF forecast data.

The ECMWF reforecast data set consists of a 15-member ensemble reforecast computed once weekly from 0000 UTC initial conditions for the initial dates of 1 September to 1 December. The years covered in the reforecast data set were from 1982 to 2001. The model cycle 29r2 was used, which was a spectral model with triangular truncation at wavenumber 255 (T255) and 40 vertical levels using a sigma coordinate system. Each forecast was run to 10 days lead. The 15 forecasts consisted of an ERA-40 reanalysis initial condition (Uppala et al. 2005) plus 14 perturbed forecasts generated using the singular-vector methodology (Molteni et al. 1996; Barkmeijer et al. 1998, 1999). While data is available to cover the entire globe, for this study the model forecasts

were extracted on a 1-degree grid from 135 to 45 degrees west longitude and 15 to 75 degrees north latitude. This covered the conterminous US and most of Canada. From this 1-degree grid, forecasts were bilinearly interpolated to the observation locations, described below.

In addition, the ECMWF 0000 UTC forecasts in the year 2005 were extracted for every day from 1 July to 1 December. This additional data permits experiments comparing short training data sets with the reforecasts. 2005 forecasts were initialized with the operational 4-dimensional variational data assimilation system (Mahfouf and Rabier 2000), rather than the 3-dimensional variational analysis of ERA-40.

b. GFS forecast data.

The GFS reforecast data set, more completely described in Hamill et al. (2006) was utilized here. It utilizes a T62, 28 sigma-level, circa-1998 version of the GFS. Fifteen-member forecasts are available to 15 days lead for every day from 1979 to current. Forecasts were started from 0000 UTC initial conditions, and forecast information was archived on a 2.5-degree global grid. GFS forecast data was also bi-linearly interpolated to surface-observation locations. For most of the experiments to be described here, the GFS reforecasts will be sub-sampled to the dates of the ECMWF reforecast data set, to permit ease of comparison. However, some experiments will utilize a 25-year (1979-2003), every-day samples of reforecast training data.

c. 2-m temperature observations.

0000 UTC and 1200 UTC 2-meter temperature observations were extracted from the National Center for Atmospheric Research (NCAR) data set DS472.0. Only observations that were within the domain of the ECMWF reforecast data set as described above were used. Additionally, only the stations that had 96 percent or more of the observations present over the 20-year period were utilized. A plot of these 439 station locations is provided in Fig. 1.

3. **Calibration and validation methodologies**.

a. Calibration with non-homogeneous Gaussian regression.

Many methods may be used for the calibration of two-meter temperature forecasts; among those in the recent literature are rank-histogram techniques (Hamill and Colucci 1998, Eckel and Walters 1998), ensemble dressing (Roulston and Smith 2003, Wang and Bishop 2005), Bayesian model averaging (Raftery et al. 2005), logistic regression (Hamill et al. 2006), analog techniques (Hamill and Whitaker 2007), and non-homogeneous Gaussian regression (Gneiting et al. 2005). Wilks and Hamill (2007) provide an inter-comparison of several of these techniques. In the inter-comparison, non-homogeneous Gaussian regression was determined to be more skillful or nearly as skillful as the other candidate techniques. Accordingly, we shall use it as the calibration technique of choice here.

Non-homogeneous Gaussian regression (NGR) is an extension to conventional linear regression. It was assumed that there may be information about the forecast uncertainty provided by the ensemble sample variance (Whitaker and Loughe 1998).

However, due to the limited number of members and other system errors, the ensemble sample variance may not properly estimate by itself the forecast uncertainty. Accordingly, the regression variance was allowed to be non-homogeneous (not the same for all values of the predictor), unlike linear regression. In this implementation of NGR, the mean forecast temperature and sample variance interpolated to the station location were predictors, and observed 2-meter temperature at station locations were the predictands. We assumed that stations had particular regional forecast biases sometimes distinct from those at nearby stations. Hence, the training did not composite the data, i.e., the fitted parameters at Atlanta were determined only from Atlanta forecasts and not from a broader sample of locations around and including Atlanta.

To describe NGR more formally, let $\sim\mathcal{N}(\alpha, \beta)$ denote that a random variable has a Gaussian distribution with mean α and variance β. Let $\overline{x}_{ens}$ denote the interpolated ensemble mean and s^2_{ens} denote the ensemble sample variance. Then NGR estimated regression coefficients a, b, c, d so as to fit $\mathcal{N}(a+b\,\overline{x}_{ens}, c+d\,s^2_{ens})$. When $d = 0$, there was no spread-error relationship in the ensemble, and the resulting distribution resembled the form of linear regression, with its constant-variance assumption. Following Gneiting et al. (2005), the four coefficients were fit iteratively to minimize the continuous ranked probability score (e.g., Wilks 2006).

In all experiments using the weekly reforecast data, cross validation was utilized in the regression analysis. The year being forecast was excluded from the training data, e.g., 1983 forecasts were trained with 1982 and 1984-2001 data. Also, because biases can change with the seasons, the full September-December data was not used as training

data. Rather, only the 5 weeks centered on the date of interest were used, e.g., when training for 15 September, the training data was comprised of 1, 8, 15, 22, and 29 September forecasts. For dates at the beginning and end of the reforecast, a non-centered training data set was used; for example, the training dates for 1 September were 1, 8, and 15 September. Unless otherwise noted, the GFS reforecast data was sub-sampled to the same weekly dates of the ECMWF training data set. However, some latter experiments will include a comparison with forecasts trained using daily GFS reforecast data from 1979-2003.

A slightly more complicated version of NGR was used for production of a calibrated multi-model ECMWF/GFS forecast. The first step was to perform a linear regression analysis of each model's ensemble-mean forecast against the observations separately for each forecast lead time. The result was an equation to predict the lowest RMS error forecast from each system's raw ensemble-mean forecast. Denote this corrected mean forecast as $\overline{x}_{EC}(k,l)$ from the ECMWF model for the kth of K training samples and lth of L locations, and similarly $\overline{x}_{GFS}(k,l)$ for the GFS. Denote the deviation of the ith of m ECMWF members from its mean as $x_{EC}^{i}{}'(k,l)$, and similarly $x_{GFS}^{i}{}'(k,l)$ for the GFS. Let D_{EC}^2 denote the average squared difference between the regression-corrected ECMWF ensemble-mean forecast and observations:

$$D_{EC}^2(l) = \frac{1}{K}\left(\overline{x}_{EC}(k,l) - o(k,l)\right)^2 \qquad , \tag{3}$$

where o(k,l) is the observation. The squared difference for the GFS, $D_{GFS}^2(l)$, is similarly defined.

We now seek to determine a multi-model weighted mean forecast and sample variance, providing larger weights to the forecasts with the smaller squared differences. The weight to apply to the ECMWF forecasts (Daley 1991, eq. 2.2.3) is defined as

$$W_{EC}(l) = \frac{D^2_{GFS}(l)}{D^2_{GFS}(l) + D^2_{EC}(l)} \quad , \tag{4}$$

and W_{GFS} = 1.0 - W_{EC} . A weighted multi-model ensemble mean was calculated as

$$\bar{x}_{MM}(k,l) = W_{EC}(l)\,\bar{x}_{EC}(k,l) + W_{GFS}(l)\,\bar{x}_{GFS}(k,l) \quad , \tag{5}$$

and a weighted multi-model ensemble variance was calculated as

$$s^2_{MM}(k,l) = W_{EC}(l)\,\frac{\sum_{i=1}^{m}\left(x^i_{EC}{}'(k,l)\right)^2}{m-1} + W_{GFS}(l)\,\frac{\sum_{i=1}^{m}\left(x^i_{GFS}{}'(k,l)\right)^2}{m-1} \quad . \tag{6}$$

These multi-model means and sample variances are then input into the NGR to produce the regression coefficients *a, b, c*, and *d*. A given forecast day's ensemble forecasts were processed using the same procedure as the training data (eqs. 4 – 6) to produce a multi-model mean and sample variance, and the regression coefficients were applied to determine the parameters of the fitted NGR distribution.

b. Validation procedures.

1) RANK HISTOGRAMS

Reliability characteristics of the probabilistic forecasts were diagnosed with rank histograms (Hamill 2001). When generating rank histograms for the "raw" unmodified, forecasts, random, normally distributed noise with a magnitude of 1.5 C was added to each member to account observation and representativeness error (ibid, section 3c). The choice of 1.5 C was somewhat arbitrary, but was generally consistent with the observation errors assigned to surface data in data assimilation schemes (Parrish and Derber 1992). Probably somewhat less random error should be added to the ECMWF forecasts than to the GFS forecasts, since the ECMWF grid spacing is smaller, lessening the representativeness error. Lacking guidance, however, the random error was set the same for both forecasts.

Rank histograms assess the rank of the observed relative to ensemble member forecasts, i.e., the observed rank is relative to discrete samples from a pdf rather than the pdf itself. How then can the rank histogram be used to assess the reliability of a fitted pdf? We used the following approach, motivated by the probability integral transform (Casella and Berger, 1990, p. 52). The original ensembles were comprised of m=15 members, so we constructed 15 sample members where the value of the ith fitted member was defined as $x_{fit}(i) = q_{i/(m+1)}$, the $i/(m+1)$th quantile of the fitted distribution. The m constructed ensemble members defined the boundaries between m+1 equally probable bins under the null hypothesis that the observed was a random draw from the same underlying distribution as the ensemble.

The $x_{fit}(i)$ was re-mapped from the i / (m+1) th quantile $q^N_{i/(m+1)}$ of a standard normal distribution. Specifically, given the coefficients *a, b, c,* and *d* that define the fitted forecast for this sample, then

$$x_{fit}(i) = q^N_{i/(m+1)}\left(c + ds^2_{ens}\right) + \left(a + b\overline{x}_{ens}\right) \tag{7}$$

(Wilks 2006, eq. 4.25). The rank of the observed relative to $x_{fit}(1) \ldots x_{fit}(m)$ was computed, and the process was repeated for all forecast samples to generate the rank histogram. Because the underlying fitted distribution was determined by training against real, imperfect observations, there was no need to perturb the ensemble members with observation noise, as was done with the raw ensemble.

2) SPREAD, ERROR, AND FRACTIONAL BIAS

Ideally, an ensemble forecast system ought to have a similar magnitude of its spread and root-mean-square (RMS) error (e.g., Whitaker and Loughe 1999). Plots of averages of these quantities are shown later, where the ECMWF's model spread at a given lead time σ_{EC} is defined as

$$\sigma_{EC} = \left\{ \frac{1}{KL} \sum_{l=1}^{L} \sum_{k=1}^{K} \left[x^{i}_{EC}{}'(k,l) \right]^2 \right\}^{1/2}, \tag{8}$$

the RMS error RMS_{EC} is defined as

$$RMS_{EC} = \left\{ \frac{1}{KL} \sum_{l=1}^{L} \sum_{k=1}^{K} \left[\overline{x}_{EC}(k,l) - o(k,l) \right]^2 \right\}^{1/2}. \tag{9}$$

Fractional bias BF_{EC} is used to diagnose how much of the ensemble-mean forecast error can be attributed to bias, as opposed to random error. It is defined as

$$BF_{EC} = \left| \frac{\sum_{l=1}^{L}\sum_{k=1}^{K}\left\{\overline{x}_{EC}(k,l) - o(k,l)\right\}}{\sum_{l=1}^{L}\sum_{k=1}^{K}\left\{\left|\overline{x}_{EC}(k,l) - o(k,l)\right|\right\}} \right| \quad . \tag{10}$$

Spread σ_{GFS}, error RMS_{GFS}, and fractional bias BF_{GFS} of the GFS forecasts are similarly defined.

3) CONTINUOUS RANKED PROBABILITY SKILL SCORE

Calculation of a revised version of the continuous ranked probability skill score (*CRPSS*) followed the method described in Hamill and Whitaker (2007). As noted in Hamill and Juras (2006), the conventional method of calculating many verification metrics, including the *CRPSS*, can provide a misleadingly optimistic assessment of the skill if the climatological uncertainty varies among the samples. The verification metric may diagnose positive skill that can be attributed to a difference in the climatologies amongst samples rather than any inherent forecast skill. Here we followed the specific method outlined in Hamill and Whitaker (2007) to ameliorate this problem. The idea was simple: divide the overall forecast sample into subgroups where the climatological uncertainty was approximately homogeneous; determine the *CRPSS* for each subgroup, and then determine the final *CRPSS* as a weighted average of the subgroups' *CRPSS*. Here, there were *NC=8* subgroups, with a more narrow range of climatological uncertainty in each subgroup, and equal numbers of samples assigned to each subgroup.

Let $\overline{CRPS}^{f}(s)$ denote the average forecast continuous ranked probability score (*CRPS*; Wilks 2006) for the *s*th subgroup, and $\overline{CRPS}^{c}(s)$ denote the average *CRPS* of the climatological reference forecast for this subgroup. Then the overall *CRPSS* was calculated as

$$CRPSS = \frac{1}{NC}\sum_{s=1}^{NC}\left(1 - \frac{\overline{CRPS}^{f}(s)}{\overline{CRPS}^{c}(s)}\right) \quad . \tag{11}$$

The climatological mean and standard deviation were calculated using 5 weeks of centered data. For more details on the calculation of the alternative formulation of the *CRPSS*, please see Hamill and Whitaker (2007).

Confidence intervals for assessing the statistical significance of differences between forecasts was done following the block bootstrap procedure outlined in Hamill (1999). 4000 iterations of a resampling procedure were used, shuffling the data in blocks of case days. The *CRPSS* was computed using eq. (10) for the two resampled sets, and the difference in *CRPSS* was used to build up the distribution for the null hypothesis. Confidence interval data will not be plotted here; for the 20-year ECMWF reforecast experiments, the 95 percent confidence intervals for calibrated vs. raw ensembles were small, from +/- 0.033 at the half-day lead to +/- 0.02 at the 10-day lead.

4. Results.

a. 20-year weekly training data.

Figure 2 provides rank histograms for the ECMWF and GFS reforecasts. For the raw forecasts the common U shape was more pronounced at the short leads and slightly more pronounced for GFS forecasts than ECMWF forecasts. After calibration with NGR, the rank histograms were much flatter, though there still was some slight excess of population of the lowest rank. Probably the assumption of Gaussianity underlying the NGR was not strictly appropriate; while perhaps forecast probability density functions (pdfs) may have somewhat more Gaussian distributions than climatology, notably 416 of the 439 stations exhibited a negative skew of their observed 2-meter temperature distributions.

The general similarity of the rank histogram shapes from the ECWMF and GFS ensembles may be somewhat misleading about the characteristics of these ensembles. Figure 3 provides a plot of average spreads (the standard deviations of the ensembles about their means; eq. 8) and the root-mean-square (RMS) errors (eq. 9) from the raw ensembles. In a perfect ensemble forecast where ensemble spread is due solely to chaotic growth of initial condition errors, these two curves should lie on top of each other. Neither the ECMWF nor the GFS ensembles had spread nearly as large as the RMS error, indicating that model biases were large. However, the RMS error of the ECMWF ensemble was substantially smaller than that of the GFS, indicating that its forecasts should have higher skill.

We now consider the overall *CRPSS* of the calibrated and uncalibrated forecasts in Fig. 4. Several main points can be made. First, as suggested by Fig. 3, the raw ECMWF forecasts were indeed more skillful than the GFS forecasts. Second, while the

raw GFS forecasts had zero or negative skill relative to climatology, after statistical correction with NGR they exceeded the *CRPSS* of the raw ECMWF forecasts, demonstrating the large skill improvement that was possible with calibration. Third, even though the ECMWF model started with substantially greater skill than the GFS, it too benefited greatly from the statistical correction. Though improvements were not as large as with the GFS, a statistically modified 4-5-day ECMWF forecast had approximately the same *CRPSS* as did the raw 1-day forecast. Fourth, consider the multi-model NGR forecast. It consistently out-performed the calibrated ECMWF forecast by a small amount, indicating that there was some independent information provided by the older, less sophisticated *GFS*. This is consistent with many previous results from the combination of information from multiple models using smaller training data sets (e.g., Vislocky and Fritsch 1995, 1997, Krishnamurti et al. 1999). Last, note that even at day 10 there is still some skill in the calibrated ECMWF and multi-model forecasts. If one considers averages over several days such as an 8-10 day average the skill increases above that of the averages of the skills at days 8, 9, and 10 (not shown). This is because some of the loss of skill is due to small errors in the timing of events.

Figure 5 demonstrates that a substantial fraction of the forecast improvement in each system can be attributed to a simple correction of model bias. The bias-corrected ensemble forecasts were generated by subtracting the mean bias (forecast minus observed) from each ensemble member in the training sample. Between 60 and 80 percent of the improvement in skill in the ECMWF forecasts can be attributed to this simple bias correction; the NGR added the remaining 20-40 percent through its regression-based correction, spread correction, and fitting of a smooth parametric

distribution. Slightly less of the improvement was attributable to bias for the GFS ensemble.

Figures 6 (a) - (c) shows the geographic distributions of day-2 skill for the raw, NGR, and bias-corrected forecasts, respectively. The raw forecasts were commonly deficient in skill in the complex terrain of the western US and Canada, presumably because the simplified terrain heights of the forecast model differed from that of the actual stations, with concomitant errors in the estimation of surface temperatures. It appeared that a simple bias correction achieved most of the impact for the stations with particularly unskillful raw forecasts. This was demonstrated in Fig. 6(d). Here, the fractional improvement of the bias correction is plotted as a function of the raw and calibrated forecasts. Letting C_{RAW} denote the *CRPSS* of the raw forecast, C_{NGR} for the calibrated forecasts, and C_{BC} for the bias-corrected forecasts, the fractional improvement *Fr* is $(C_{BC} - C_{RAW}) / (C_{NGR} - C_{RAW})$. Figure 6(d) shows several interesting characteristics. First, note that the effect of the NGR calibration was primarily to improve forecasts that started off particularly unskillful, homogenizing the resultant skill relative to the highly varying skills seen in the raw forecasts. Second, in general the locations that had relatively large improvements through the NGR calibration achieved a greater fraction of this from the bias correction than did the locations that had smaller improvements. Overall, the large improvements from bias corrections may indicate that additional resolution may be helpful, leading to smaller mismatches between model terrain height and station elevation (see also Buizza et al. 2007).

b. Differences between 20-year weekly and 30-day daily training data sets.

To facilitate a comparison of long and short training data sets, the ECMWF and GFS ensemble forecasts were also extracted every day for the period 1 September – 1 December 2005. This permitted us to examine the efficacy of a smaller training data set. Recent results (Stensrud and Yussouf 2005, Cui et al. 2006) have suggested that temperature forecast calibration may be able to be performed well even with a small number of recent forecasts. This may be because the ensemble forecast bias is relatively consistent and can be estimated with a small sample. Another possibility is that recent samples are more relevant for the statistical correction, with their more similar circulation regimes and land-surface states than data from other years.

Accordingly, we compared the calibration of forecasts using the prior 30 days as training data to calibration using the full reforecast training data set. Forecasts were compared for the period of 1 October – 1 December 2005. Non-homogeneous Gaussian regression was again used for the calibration. Figure 7 shows that at short forecast leads, the 30-day training data set provided approximately equal skill improvements relative to the 20-year training data set for the ECMWF model, and marginally less for the GFS. However, as the forecast lead increased, then the benefit of the longer training data set becomes apparent.

Why were more samples particularly helpful for the longer leads? We suggest that there were at least three contributing factors. First, the prior 30-day training data set was 9 days older for a 10-day forecast (training days -39 to -10) than for a 1-day forecast (training days -30 to -1). If errors were synoptically dependent and a regime change took place in the intervening 9 days, the training set at 1-day lead will have had samples from

the new regime while the training set at 10-days lead will not. The second reason is that determining the bias to a pre-specified tolerance will require more samples at the long leads than at the short leads. At these long leads, the proportion of the error attributable to bias shrinks due to the rapid increase of errors due to chaotic error growth. This is shown in Fig. 8; for the ECMWF model, this decreased from ~0.54 at the half-day lead to ~0.28 at the 10-day lead. Consequently, as the overall error grows as forecast lead increases and a larger proportion of it is attributable to random errors, determining the bias to a prespecified tolerance will require more samples, by central-limit theorem arguments. The third reason was that the short-lead forecast training data sets were comprised of samples that tended to have more independent errors than the longer-lead training data sets. The ECMWF 1-day lagged correlation of forecast minus observed averaged over all stations (not shown) increased from around 0.2 at the early leads to 0.5 at the longer leads. Using the definition of an effective sample size *n*' (Wilks 2006, p. 144)

$$n' = n\frac{1-\rho_1}{1+\rho_1}, \qquad (12)$$

with *n*=30, this indicated that the effective sample size was approximately 20 at the short leads and 10 at the longer leads. The once-weekly, 20-year reforecast data set should, in comparison, be comprised of samples that were truly independent of each other.

Considering again the puzzling result of similar skill at short leads, we hypothesize that the two factors here may have contributed to underestimating the potential skill that can be obtained with a properly constructed long training data set.

First, one limitation of the ECMWF data sets was that for the 2005 data, all forecasts were initialized 4D-Var, while the 1982-2001 reforecast data were initialized with 3D-Var. It is thus possible that the ECMWF short-term reforecasts may have subtly different biases than the 2005 real-time forecasts, differences that diminish with the forecast lead. This would affect the calibration of the short-term forecasts. Notice that Fig. 7b shows a somewhat larger benefit from long training data sets with the GFS, where a consistent data assimilation system was used. Second, the calibration with the full reforecast training data set here used only the model-forecast temperature as a predictor. Perhaps the short training data set benefits from having samples with a more similar set of land-surface conditions. If this were the case, then perhaps a multi-predictor regression analysis including, say, soil-moisture content as an additional predictor would improve the reforecast calibrations.

c. Differences between 20-year weekly and 25-year daily training data sets in the GFS.

Figure 9 shows the CRPSS of GFS forecasts from the raw ensemble, after a calibration with the 20-year weekly training data set, and with the full 25-year daily training data set. When training with the 25-year daily data, training data was used in a window of +/- 15 days around the date being forecast, e.g., 16 September forecasts used 1 September – 1 October reforecasts for training data. Data for the year being forecast was excluded (cross validation). As Fig. 9 shows, there is a small but consistent difference the 25-year, daily training data set provides a small but consistent improvement over the 20-year, weekly training data set. Why isn't the improvement larger? First, of course, the baseline for the comparison used 20 years of weekly forecasts * 5 weeks of centered

data = 100 samples, a respectably large number for the estimation of the four NGR parameters. Further, as noted in Hamill et al. (2004), forecast errors may be correlated from one day to the next, so using daily vs. weekly samples does not necessarily mean that the effective sample size (Wilks 2006, p. 144) will be seven times larger with daily samples.

5. **Conclusions**

A prior series of articles (Hamill et al 2004, 2006, Hamill and Whitaker 2006, 2007, Whitaker et al. 2006, Wilks and Hamill 2007) have discussed the benefit of calibrating probabilistic forecasts using the large training data sets from an ensemble reforecast data set from a 1998 version of the NCEP GFS. This data set is now 10 years old, and it is not clear whether the large positive benefits from the large training data set would still occur with a newer, higher-resolution model with its reduced systematic errors. Recently, ECMWF developed a limited reforecast data set consisting of a once-weekly, 15-member reforecast for the period 1 September – 1 December 1982-2001. These forecasts were conducted using the model version operational in the second half of 2005, a T255-resolution version of the forecast model. While the once-weekly reforecasts were more sparse than the daily reforecasts from the GFS, the ECMWF reforecast data set still spanned two decades of diverse climatological regimes. Accordingly, we performed an analysis of the skill that can be gained from calibration of surface temperatures using these training data sets.

Both the ECMWF and GFS raw ensemble surface temperature forecasts were found to be biased and/or under-dispersive, noted by the excess populations of the

extreme ranks in their rank histograms. This tendency was more pronounced at the short forecast leads. However, after calibration with non-homogeneous Gaussian regression (NGR), the rank histograms were flatter, though the lowest rank was still populated slightly more than was appropriate with a perfectly calibrated ensemble.

The skill of these forecasts was measured with a modified version of the continuous ranked probability skill score (CRPSS), with the computation adjusted to remove the tendency to award fictitious skill due to variations in the forecast climatology (Hamill and Juras 2006). Climatology provided the no-skill reference. Using this skill measure, the raw GFS ensemble forecasts had near zero to negative skill at all leads due to the presence of large forecast biases. The ECMWF raw forecasts retained positive skill to approximately 8 days.

After calibration with NGR, the post-processed GFS forecasts exceeded the skill of the uncalibrated ECMWF forecasts at all leads. Here, the GFS training data was sub-sampled to the same weekly, 20-year set of dates as in the ECMWF reforecast. However, the reforecast-based, calibrated ECMWF forecasts were much more skillful than both the GFS calibrated forecasts and the ECMWF uncalibrated forecasts, though the absolute amount of skill increase from calibration was smaller for ECMWF than for the GFS. Nonetheless, the ECMWF skill improvement was substantial; for example, the skill of a calibrated, 4-5 day ECMWF forecast was comparable to the skill of an uncalibrated 1-day forecast. Approximately 70 percent of the improvement of the ECMWF could be attributed to a simple correction of mean bias in the forecasts, with a slightly smaller percentage in the GFS. The ECMWF raw forecasts were observed to

have particularly low skill at stations in the inter-mountain western US, perhaps due to larger mismatches between the model terrain and the station locations. Calibration was particularly successful in increasing the skill at these stations. Finally, a multi-model calibrated forecast was more skillful than either individual calibrated forecast.

The computation of an extensive reforecast data set is expensive, and a new reforecast data set may be needed each time a model change affects its systematic error characteristics. If the same benefit could be achieved with a much smaller set of recent forecasts, this would make operational calibration much easier. Accordingly, using 2005 data, we compared the calibration using the 1982-2001 reforecasts to calibration using the most recent 30 samples of forecasts from 2005. For the shorter forecast leads, the skill after calibration using this shorter training data set was very similar to that achieved with large reforecast data set. We hypothesize that this benefit may be attributable to the more recent samples being more similar in their error characteristics than those from the reforecast data set, which samples other different years of data. However, at longer leads, the reforecast data set produced more skillful calibrated forecasts than the 30-day training data set. This was likely due to at least three reasons: first, 30 days of training data for the longer-lead forecasts were more separated from the actual forecast day of interest (e.g., calibrating a 10-day forecast, the most recent training sample is 10 days old, since verification is not yet available for the more recent forecasts). Second, the number of samples necessary to estimate the bias to a prespecified tolerance generally increased with increasing forecast lead. And third, for forecasts at the longer leads, the samples on adjacent days tended to have correlated forecast errors, reducing the effective sample size.

While a daily reforecast data set was yet not available for the ECMWF model, the impact of daily vs. weekly samples could be evaluated with the GFS reforecast data set. Using a 25-year, daily reforecast vs. a 20-year weekly forecast produced a small but noticeable improvement.

It is also possible that the calibration could be improved by including other predictors. Here we considered only 2-meter temperature as a predictor. Perhaps the reason the 30-day training data set shows such good results is that the training samples are from a regime with similar surface characteristics, such as soil moisture. If so, then the performance of a multi-year reforecast could be enhanced by including soil moisture as an additional predictor. An examination of the potential value of several other predictors may be useful before any operational implementation of a temperature-calibration scheme.

This article considered only the calibration of 2-meter temperature forecasts. Our experience with precipitation calibration using the GFS reforecasts suggests that the benefit from calibration using short training data sets will be smaller than for temperature. The companion article, Part II (Hamill et al. 2007) examines the calibration of ECMWF and GFS precipitation forecasts in more depth and provides substantial further evidence for the value of large training data sets, even with a state-of-the-art model. Nonetheless, the value of large training data sets for temperature calibration was confirmed here, even for a current, state-of-the-art forecast model. Short training data sets were adequate for the short-lead forecasts, but in order to achieve benefits at all forecast leads, the longer training data set proved useful.

Combined with the evidence in Part II and previous studies, there is now a growing body of literature indicating the potential utility of reforecast methodology for improving operational ensemble predictions.

References

Barkmeijer, J., M. van Gijzen, and F. Bouttier, 1998: Singular vectors and estimates of the analysis error covariance metric. *Quart. J. Royal Meteor. Soc.*, **124**, 1695-1713.

----------, R. Buizza, and T. N. Palmer, 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Quart. J. Royal Meteor. Soc.*, **125**, 2333-2351.

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076-1097.

Buizza, R., Bidlot, J.-R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G., & Vitart, F., 2007: The new ECMWF VAREPS. *Quart. J. Royal Meteor. Soc.*, **133**, 681-695.

Casella, G., and R. L. Berger, 1990: *Statistical Inference*. Duxbury Press, 650 pp.

Clark, M.P. and L.E. Hay (2004): Use of medium-range weather forecasts to produce predictions of streamflow. *J. Hydrometeor.,* **5**, 15-32.

Cui, B., Toth, Z., Zhu, Y., Hou, D., Unger, D., Beauregard, S., 2006: The trade-off in bias correction between using the latest analysis/modeling system with a short, versus an older system with a long archive. *Proceedings, First THORPEX International Science Symposium*. December 6-10, 2004, Montréal, Canada, World Meteorological Organization, 281-284.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, 457 pp.

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting*, **13**, 1132-1147.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble Model Output Statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098-1118.

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711-724.

----------, 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155-167.

----------, 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550-560.

----------, J. S. Whitaker, and X. Wei, 2004: Ensemble re-forecasting: improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434-1447.

----------, ----------, and S. L. Mullen, 2006: Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33-46.

----------, and ----------, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev.*, **134**, 3209-3229.

----------, and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quart. J. Royal Meteor. Soc.*, **132**, 2905-2923.

----------, and J. S. Whitaker, 2007: Ensemble calibration of 500 hPa geopotential height and 850 hPa and 2-meter temperatures using reforecasts. *Mon. Wea. Rev.*, **135**, 3273-3280. .

----------, R. Hagedorn, and J. S. Whitaker, 2007: Probabilistic forecast calibration

using ECMWF and GFS ensemble reforecasts. Part II: precipitation. *Mon. Wea. Rev.*, submitted. Available at www.cdc.noaa.gov/people/tom.hamill/ecmwf_refcst_ppn.pdf .

Krishnamurti, T. N., and co-authors, 1999: Improved weather and seasonal climate forecasts from multi-model superensemble. *Science*, **285**, 1548-1550.

Mahfouf, J.-F., and F. Rabier, 2000: The ECMWF operational implementation of four-dimensional variational assimilation. II: Experimental results with improved physics. *Quart. J. Royal Meteor. Soc.*, **126**, 1171-1190.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Royal Meteor. Soc.*, **122**, 73-199.

Parrish, D. F. and J. C. Derber, 1992: The National Meteorological Center's spectral statistical interpolation analysis system. *Mon. Wea. Rev.*, **120**, 1747-1763.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155-1174.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16-30.

Stensrud, D.J., and N. Yussouf, 2005: Bias-corrected short-range ensemble forecasts of near surface variables. *Meteor. Appl.*, **12**, 217-230.

Uppala, S. M., and coauthors, 2005: The ERA-40 re-analysis. *Quart. J. Royal Meteor. Soc.*, **131**, 2961-3012.

Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bull. Amer. Meteor. Soc*., **76**, 1157-1164.

---------- , and ---------- , 1997: Performance of an advanced MOS system in the 1996-97 National Collegiate Weather Forecasting Contest. *Bull. Amer. Meteor. Soc*., **78**, 2851-2857.

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Royal. Meteor. Soc*., **131**, 965-986.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble-mean skill. *Mon. Wea. Rev*., **126**, 3292-3302.

---------- , X. Wei, and F. Vitart, 2006: Improving week two forecasts with multi-model reforecast ensembles. *Mon. Wea. Rev*., **134**, 2279-2284.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*, 2nd Ed., Academic Press, 627 pp.

---------- , and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379-2390.

FIGURE CAPTIONS

Figure 1: Station locations where probabilistic 2-meter temperature forecasts are evaluated.

Figure 2: Rank histograms for 2-meter temperatures from ECWMF and GFS ensembles at 1, 4, and 7 days lead. Histograms denote the raw ensemble and solid lines the calibrated ensembles.

Figure 3. Average ensemble spread and root-mean-square error of 2-meter temperature forecasts from (a) ECMWF ensemble and (b) GFS ensemble.

Figure 4: CRPSS of surface temperature forecasts with and without calibration.

Figure 5: CRPSS including bias-corrected ensemble forecasts for (a) ECMWF forecasts, and (b) GFS forecasts.

Figure 6: (a) *CRPSS* of raw 2-day forecasts from the ECMWF model. (b) as in (a), but for calibrated NGR forecasts, (c) as in (a) but for bias-corrected forecasts. (d) Fractional improvement *Fr* gained from bias correction as a function of the *CRPSS* from raw and NGR forecasts.

Figure 7: Comparison of CRPSS using 30-day and 20-year training data sets for the period 1 October – 1 December 2005. (a) ECMWF data, (b) GFS data.

Figure 8: Fractional bias, the fraction of the total RMS error that can be attributed to systematic error, as a function of forecast lead.

Figure 9: CRPSS of GFS forecasts from raw ensemble, with 20-year weekly training data set, and 28-year daily training data set.

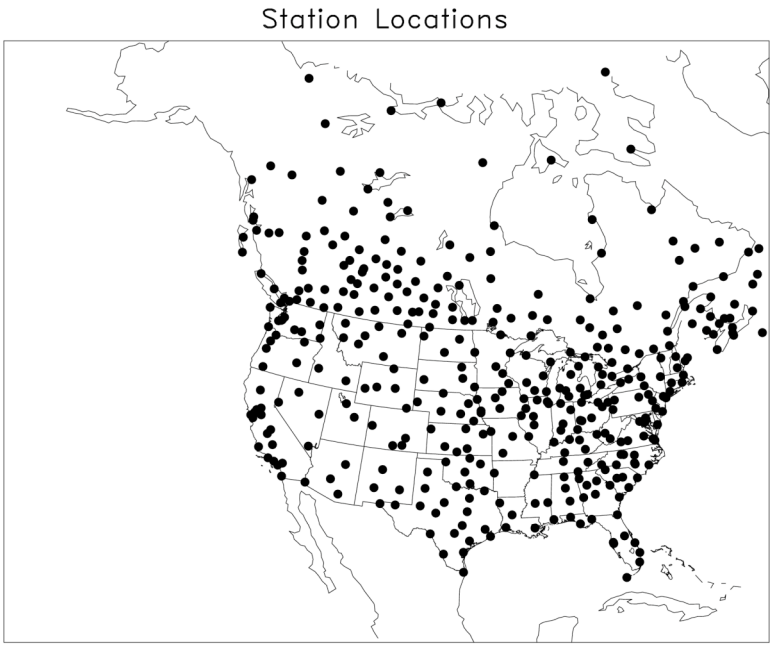

Figure 1: Station locations where probabilistic 2-meter temperature forecasts are evaluated.

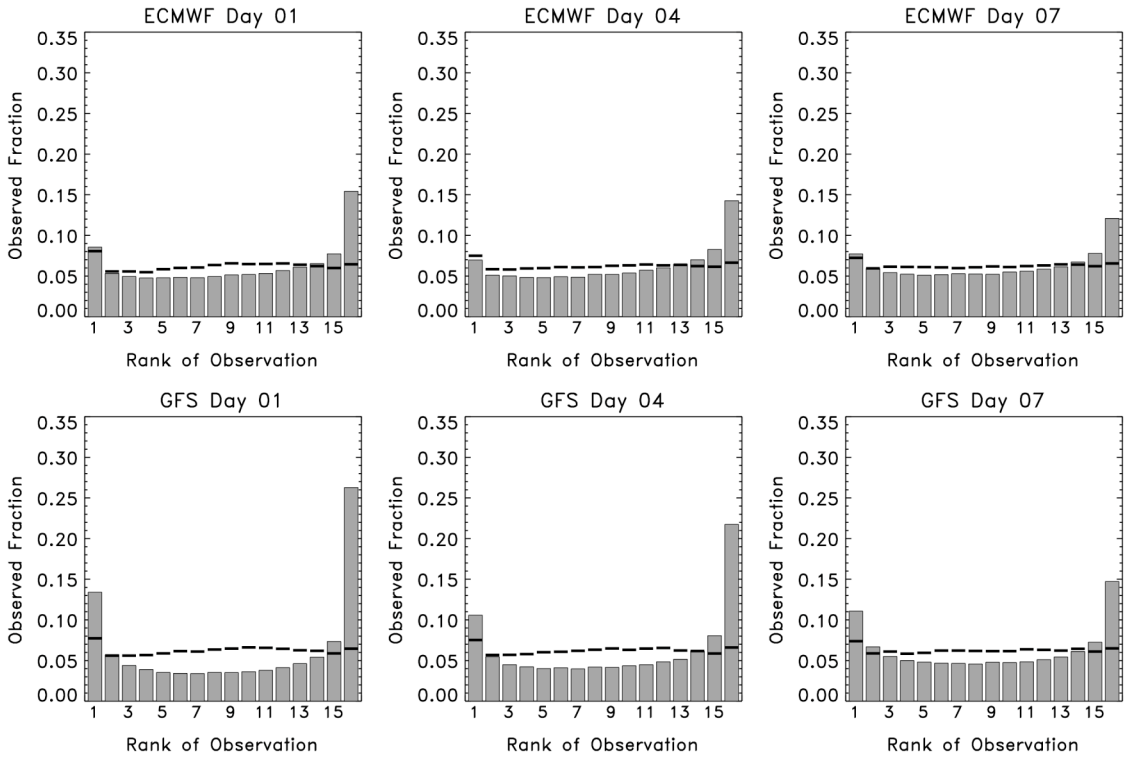

Figure 2: Rank histograms for 2-meter temperatures from ECWMF and GFS ensembles at 1, 4, and 7 days lead. Histograms denote the raw ensemble and solid lines the calibrated ensembles.

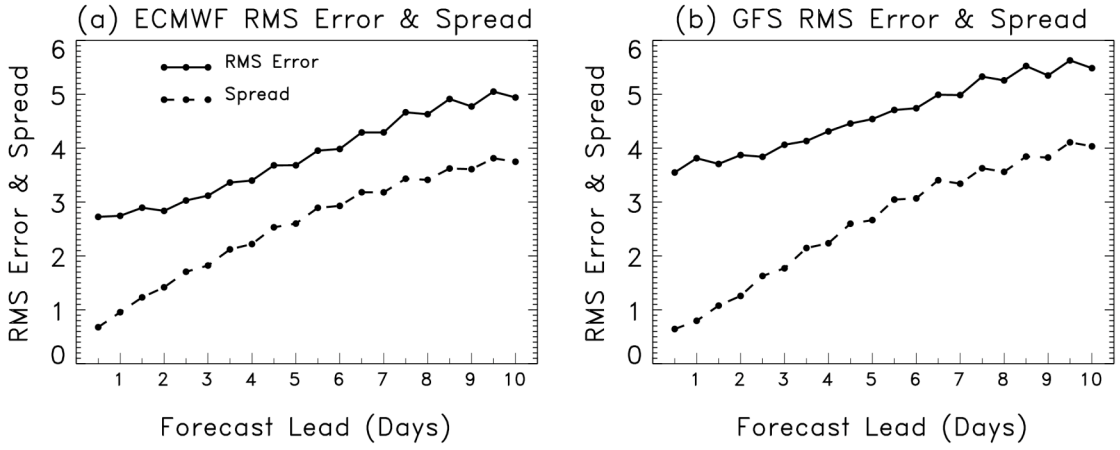

Figure 3. Average ensemble spread and root-mean-square error of 2-meter temperature forecasts from (a) ECMWF ensemble and (b) GFS ensemble.

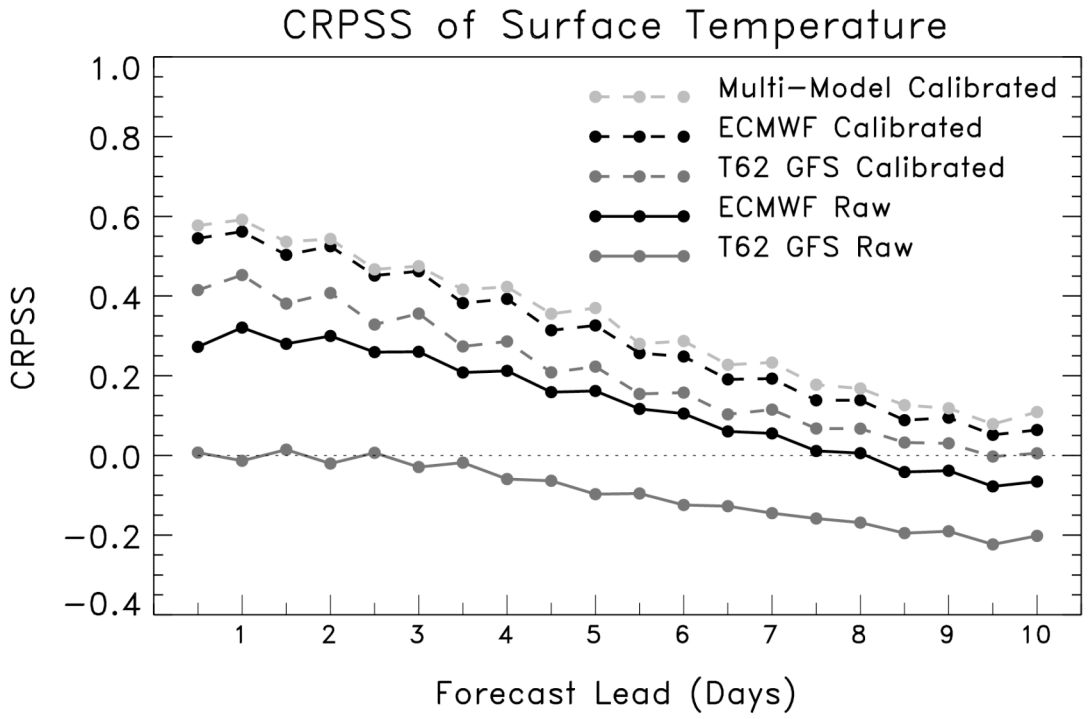

Figure 4: CRPSS of surface temperature forecasts with and without calibration.

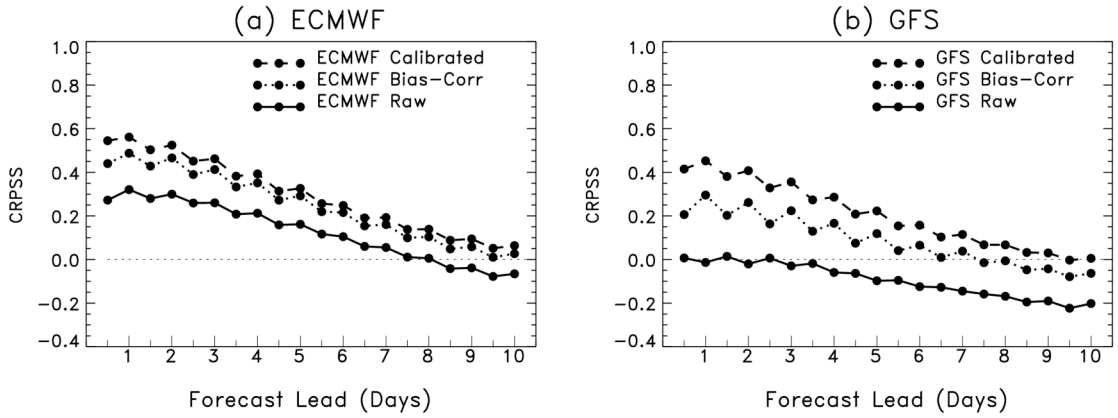

Figure 5: CRPSS including bias-corrected ensemble forecasts for (a) ECMWF forecasts, and (b) GFS forecasts.

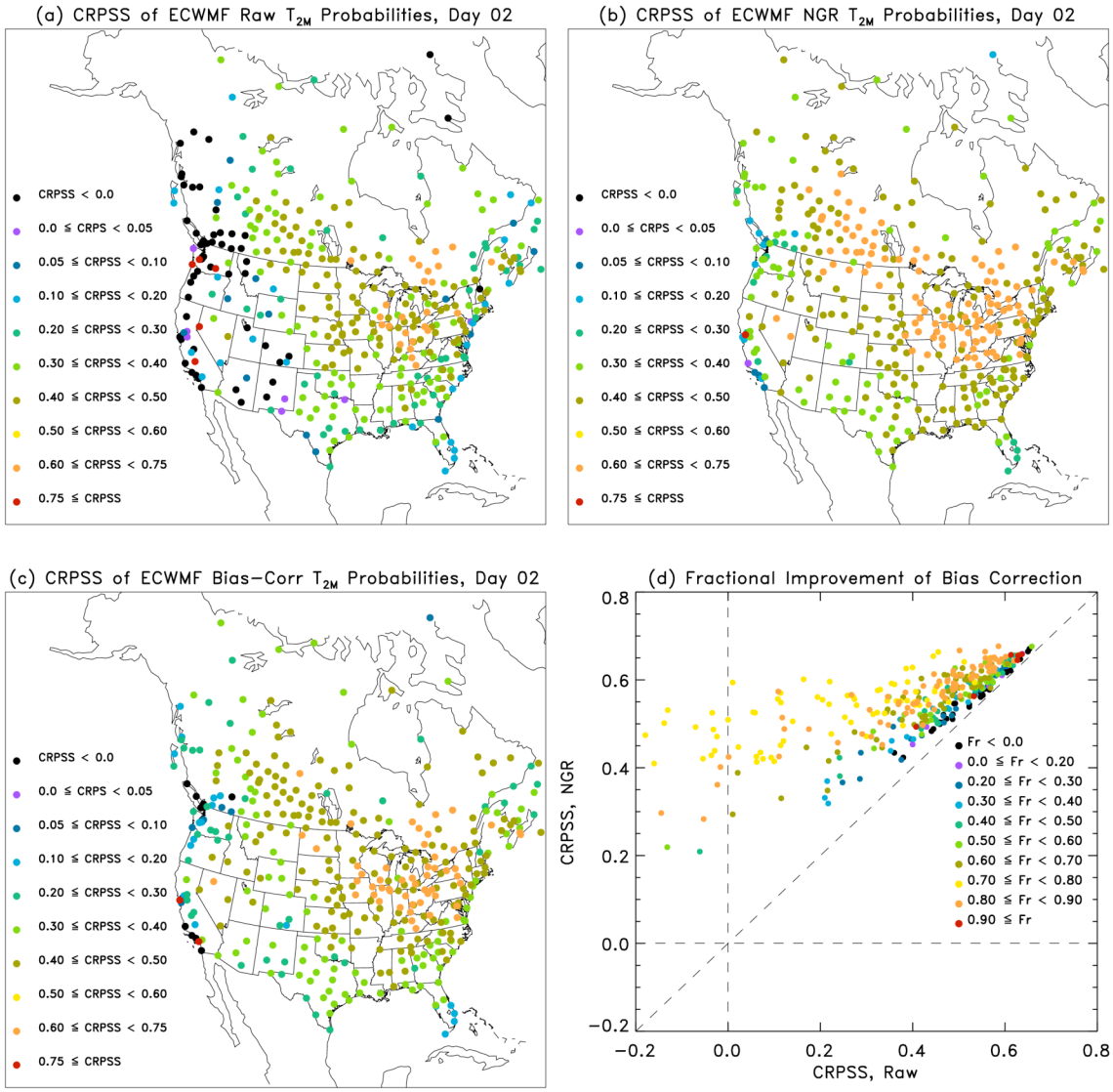

Figure 6: (a) *CRPSS* of raw 2-day forecasts from the ECMWF model. (b) as in (a), but for calibrated NGR forecasts, (c) as in (a) but for bias-corrected forecasts. (d) Fractional improvement *Fr* gained from bias correction as a function of the *CRPSS* from raw and NGR forecasts.

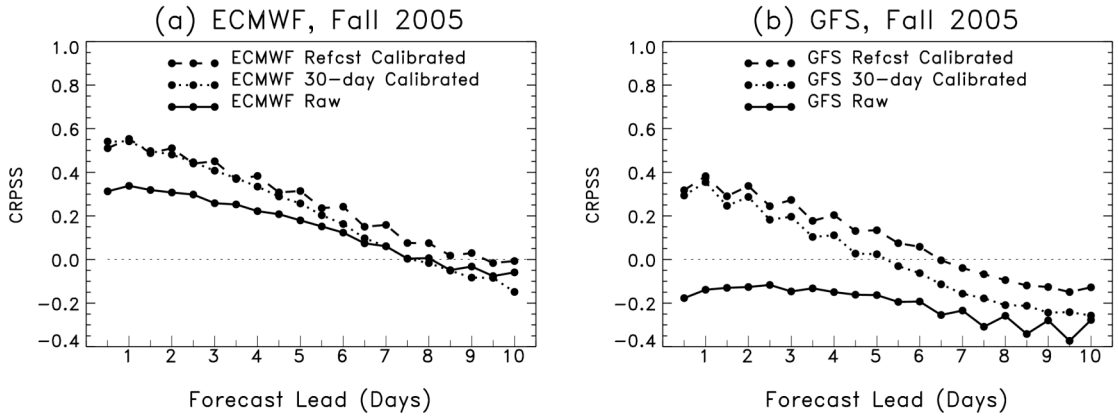

Figure 7: Comparison of CRPSS using 30-day and 20-year training data sets for the period 1 October – 1 December 2005. (a) ECMWF data, (b) GFS data.

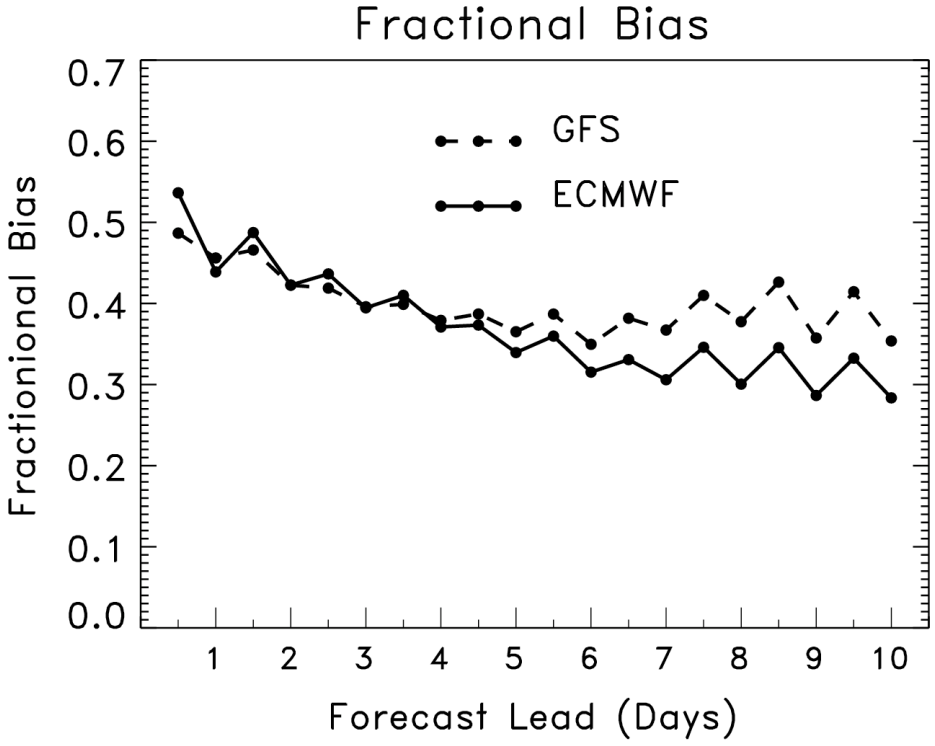

Figure 8: Fractional bias, the fraction of the total RMS error that can be attributed to systematic error, as a function of forecast lead.

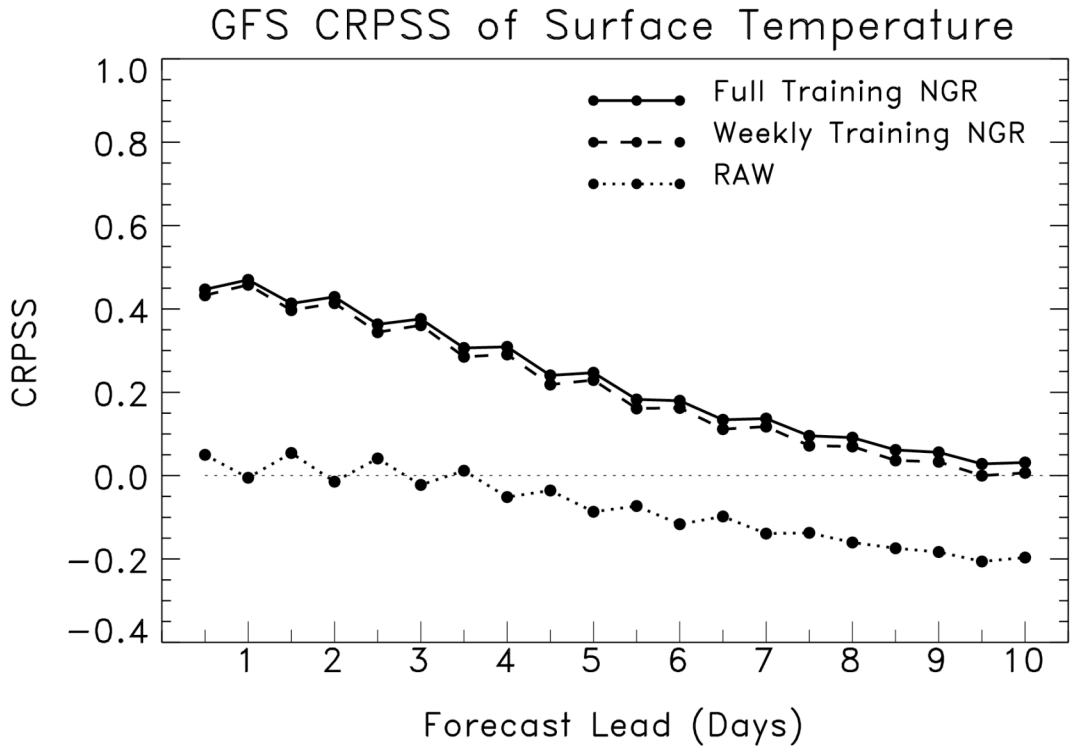

Figure 9: CRPSS of GFS forecasts from raw ensemble, with 20-year weekly training data set, and 28-year daily training data set.